ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

ПРЕДСКАЗАНИЕ ДОЛИ ИОНА ПЕПТИДА ЗАДАННОГО ЗАРЯДА В МАСС-СПЕКТРОМЕТРИИ С ПОЛОЖИТЕЛЬНОЙ ЭЛЕКТРОСПРЕЙНОЙ ИОНИЗАЦИЕЙ

В.С. Скворцов*, Н.Н. Алексейчук, Ю.В. Мирошниченко, А.В. Рыбина

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича
119121, Россия, Москва, ул. Погодинская, 10; *e-mail: vladlen@ibmh.msk.su

Рассмотрена возможность предсказания по аминокислотной последовательности доли иона заданного заряда в общем количестве пептида при масс-спектрометрии с положительной ионизацией электроспреем. В качестве исходных использовали данные, полученные в эксперименте MS/MS [Ramus et al., 2015, Data in brief, 6, 286-294] с использованием стандартизованного набора UPS1 (48 высокоочищенных белков человека) и депонированные в ProteomeXchange (идентификатор PXD001819). Для каждого из идентифицированных пептидов, принадлежащих одному из белков набора UPS, формировали список из обнаруженных ионов различного заряда. Сумма интенсивностей пиков, обнаруженных для первичного иона, служила в качестве меры количества пептида. Так как соотношения долей пептида между ионами разного заряда не зависит от концентрации в экспериментальной пробе, общая выборка была сформирована объединением данных, полученных для различных разведений UPS1. Построен набор уравнений, с помощью которых показана возможность предсказать долю ионов 1+, 2+ и 3+.

Ключевые слова: пептид; масс-спектрометрия; электроспрей; предсказание свойств

DOI: 10.18097/BMCRM00100

ВВЕДЕНИЕ

На настоящий момент масс-спектрометрия (MS) является основным инструментом протеомных исследований [1]. Среди различных вариантов ионизации широко используется такой метод мягкой ионизации, как ионизация электроспреем (ESI). При анализе пептидов, например при тандемной масс-спектрометрии (MS/MS), количество ионов определённого заряда зависит от используемого оборудования, условий эксперимента (приложенное напряжение, концентрация и скорость потока раствора и т.д.), состава растворителя [2]. В то же время очевидно, что распределение ионов при ESI в одних и тех же условиях определяется в первую очередь аминокислотной последовательностью пептида. Таким образом, при планировании эксперимента (например, при выборе способа гидролиза или выборе рабочего окна для регистрации ионов с заданными величинами *m/z*, при выборе условий отслеживания ионов с определённым зарядом) важно знать, какие зарядные состояния и в какой доле от общего количества пептида можно будет зарегистрировать. Знать долю от общего количества пептида важно также в экспериментах по количественному определению белка с помощью масс-спектрометрических исследований. Конечно, это возможно только при целом ряде допущений: количество различных пептидов при гидролизе примерно совпадет и пропорционально общему количеству белка; «невидимая часть» или нерегистрируемая доля пептидов примерно одинаковая и т.д. Однако представляется, что даже приблизительное предсказание с точностью до 10-15% может качественно улучшить данные масс-спектрометрического эксперимента. В данной работе исследуется возможность предсказания доли иона с конкретным зарядом на основе аминокислотного спектра пептида (последовательности пептида).

МЕТОДИКА

Наиболее важным фактором в подобного рода работе является наличие качественных, а ещё лучше, и стандартизированных данных. В работе использованы данные, полученные в работе [3], депонированные в ProteomeXchange [4] (идентификатор PXD001819). Данные получены в ходе масс-спектрометрического исследования с использованием стандартизованного набора из 48 высокоочищенных белков человека без SAP (PTM возможны), как синтезированных рекомбинантно, так и полученных из естественных источников и представленных в пробе в одинаковой концентрации (UPS1 компании «Sigma-Aldrich», США). К особенностям масс-спектрометрического эксперимента относится следующее[3]:

1. пробы для идентификации авторы готовили, смешивая лизат дрожжевых клеток с набором разведений UPS таким образом, чтобы получить конечную концентрацию UPS1 0.05 fmol/mg, 0.125 fmol/mg, 0.250 fmol/mg, 0.5 fmol/mg, 2.5 fmol/mg, 5 fmol/mg, 12.5 fmol/mg, 25 fmol/mg и 50 fmol/mg дрожжевого лизата;
2. гидролиз проводили в растворе с добавлением 2% трипсина;
3. nanoLC-MS/MS проводили с использованием системы nanoRS UHPLC («Dionex», Нидерланды, совмещенной с

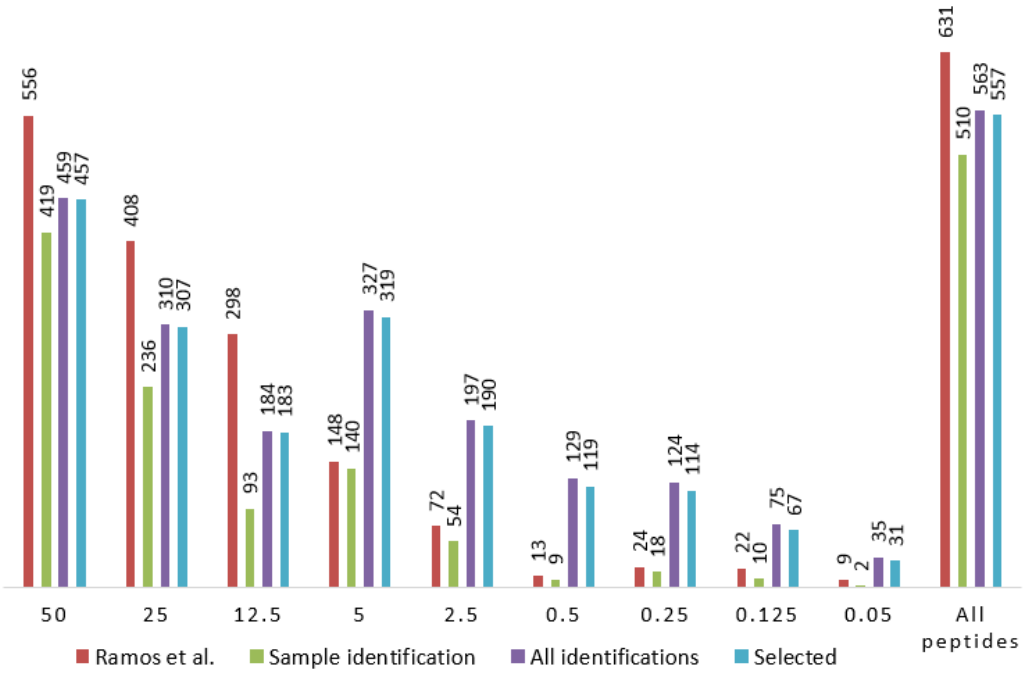

Рисунок 1. Распределение количества идентифицированных пептидов в каждом из разведений, полученных в работе [3], и при объединении данных в сводную выборку. Ramos et al. - данные из работы [3]. Sample identification – использованы пептиды, идентифицированные для конкретного разведения. All identifications – использованы все пептиды, идентифицированные в работе [3], вне зависимости от разведения. Selected – дополнительный отбор (точность при сравнении первичных ионов 5 ppm, удаление пептидов с N-концевым ацетилированием).

масс-спектрометром LTQ-Orbitrap Velos («Thermo Fisher Scientific», Германия) по 3 повтора для каждого разведения;
4. идентификацию пептидов авторы проводили двумя способами: с использованием Mascot Daemon версии 2.4 («Matrix Science», Великобритания) с точностью 5 ppm для первичного иона и 0.8 Da для идентификации фрагментов; с использованием комбинации программ MaxQuant [5] и поисковой машины Andromeda [6] (6 ppm и 0.5 Da).

В ходе анализа мы совместили данные по идентификации пептидов с данными по распределению первичных ионов различной зарядности (включая суммарную интенсивность всех пиков для данного иона (S_i) как меру количества детектированных ионов). Данные по всем зарегистрированным первичным ионам были экстрагированы нами из raw-файлов с помощью программы Dinosaur [7], которая кроме суммы интенсивностей всех пиков данного иона вычисляет также начальное время появления и исчезновения данного иона на выходе с хромотографической колонки и время, соответствующее максимальной интенсивности при детекции (время удержания или RT). Затем данные группировали следующим образом. Группы первичных ионов с различным зарядом считали принадлежащим одному пептиду, если масса полного пептида находилась в пределах 5 ppm, времена выхода с колонки перекрывались, а положение максимума отличалось не более чем на 0.2 мин. При сравнении с пептидами, идентифицированными авторами в работе [3], считали, что первичный ион соответствовал идентифицированному пептиду, если значения *m/z* одинаково заряженных ионов совпадали в пределах 5 ppm, и RT иона, измеренного авторами работы [3], лежало в пределах диапазона RT, вычисленного программой Dinosaur. При формировании выборки распределения первичных ионов по заряду данные усредняли по всем приведенным повторам. При поиске соответствия между данными о распределении первичных ионов и данными по идентификации MS/MS, последние использовались без привязки к конкретному разведению (см. выше п.1). При наличии модификаций данные с N-концевым ацетилированием в выборку не включали, при наличии остатка окисленного метионина использовали только вариант с максимальной суммарной интенсивностью. Данные, полученные в результате такого анализа, представлены в сводной таблице дополнительных материалов.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

На рисунке 1 представлено распределение числа идентификаций пептидов в каждом из разведений, полученных в работе [3], и при объединении данных в сводную таблицу (дополнительные материалы) по правилам, описанным в разделе «Методика». Видно, что при использовании пептидов, идентифицированных только для соответствующего разведения, существенная часть данных теряется. В общей сложности не найдено соответствия для 121 пептида из 631, идентифицированных в работе [3] (в пределах отдельных разведений различия ещё больше). Из них 69 были обнаружены, но не удовлетворяли установленным параметрам соответствия по точности (5 ppm). Ещё 15 надо было исключить из анализа как содержащих запрещённую модификацию (N-концевое ацетилирование). Учитывая, что во всех опытах условия экспериментов отличались только концентрацией белков UPS1, мы полагаем, что использовать данные о заряде, *m/z* и RT иона, идентифицированного для

Таблица 1. Параметры уравнений линейной регрессии, предсказывающих значения величин ω_{n+} $\log(C_{n/m})$, полученные при обучении и в процедуре скользящего контроля.

№ выборки	Описание выборки	Зависимая величина	n	R^2_L	SEM_L	Q^2	SEM_{LOO}
1.	Вар.1	ω_{1+}	557	0.2	0.031	-	-
2.	Вар.1	ω_{2+}	557	0.53	0.208	0.5	0.214
3.	Вар.1	ω_{3+}	557	0.35	0.229	-	-
4.	Вар.1	ω_{4+}	557	0.29	0.08	-	-
5.	Вар.2	ω_{1+}	148	0.45	0.44	0.3	0.051
6.	Вар.2	ω_{2+}	300	0.64	0.153	0.59	0.162
7.	Вар.2	ω_{3+}	177	0.47	0.107	0.29	0.130
8.	Вар.2	ω_{4+}	29	-	-	-	-
9.		$\log(C_{1/2})$	148	0.65	0.183	0.51	0.216
10.		$\log(C_{2/3})$	161	0.55	0.384	0.45	0.42
11.		$\log(C_{3/4})$	25	-	-	-	-

Примечание. R^2_L – R^2 обучения; SEM_L – средняя ошибка обучения; n – число наблюдений при обучении; Q^2 - Q^2 модели; SEM_{LOO} – среднеквадратичная ошибка для метода скользящего контроля; R^2_T – R^2 предсказания для тестовых выборок (номер выборки указан в скобках). Вар.1 - с учётом 0 и 1. Вар.2 - Без учёта 0 и 1. Вычисления не выполняли, если число наблюдений было меньше 60, либо R^2 обучения был меньше 0.4.

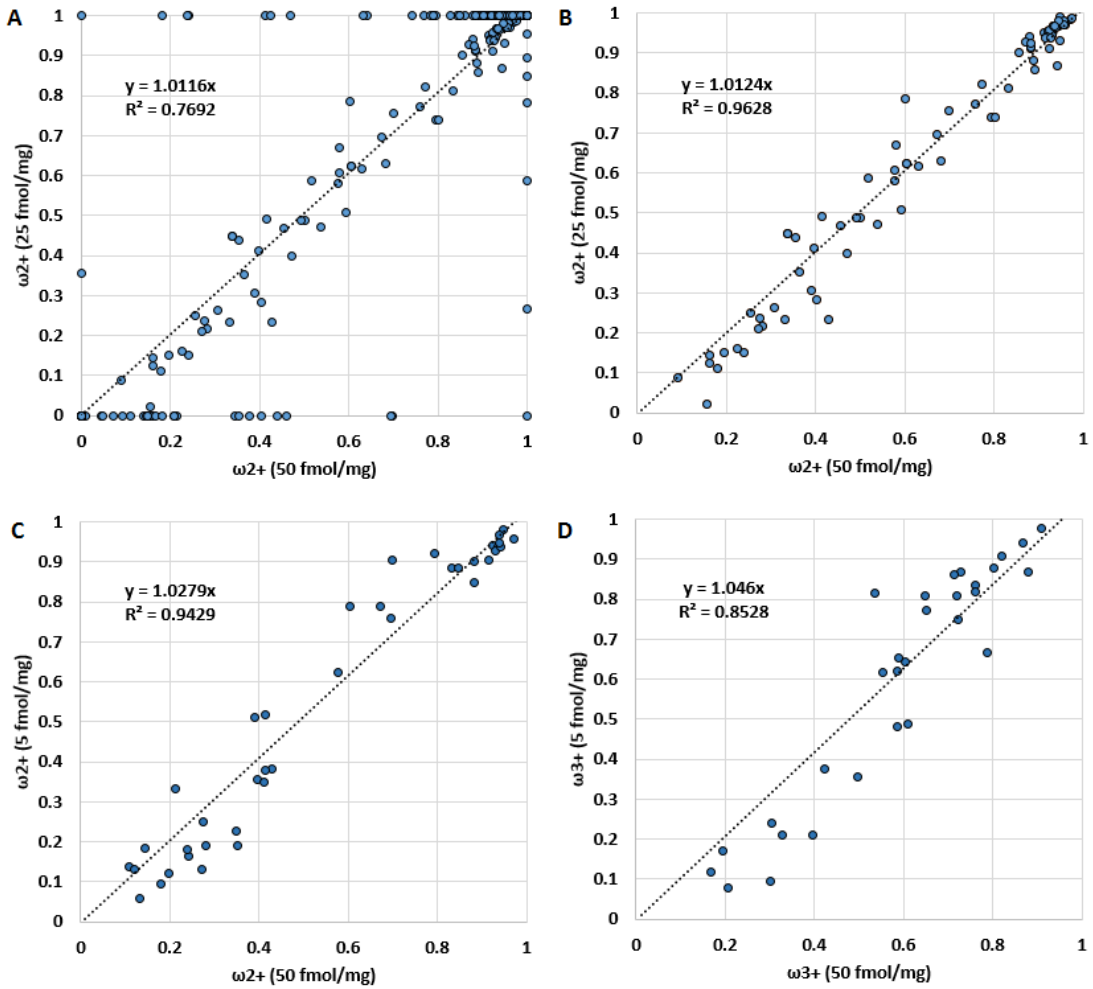

Рисунок 2. Примеры сравнения доли иона пептида конкретной зарядности (ω_{n+}), полученной для различных разведений. Вариант А включает пептиды, для которых в данных разведениях величина ω_{n+} равна 1 (обнаружен только один вариант иона) и 0 (ион с данным зарядом не зарегистрирован)

другого разведения, правомерно для всех, так как «выбор иона» для фрагментации и получения качественного для идентификации спектра, строго говоря, не детерминирован и может носить случайный характер.

Поскольку суммарная интенсивность – величина, зависимая от настроек прибора и случайных факторов, меняющихся от опыта к опыту, то в качестве меры количества иона с заданным зарядом использовали 2 набора величин.

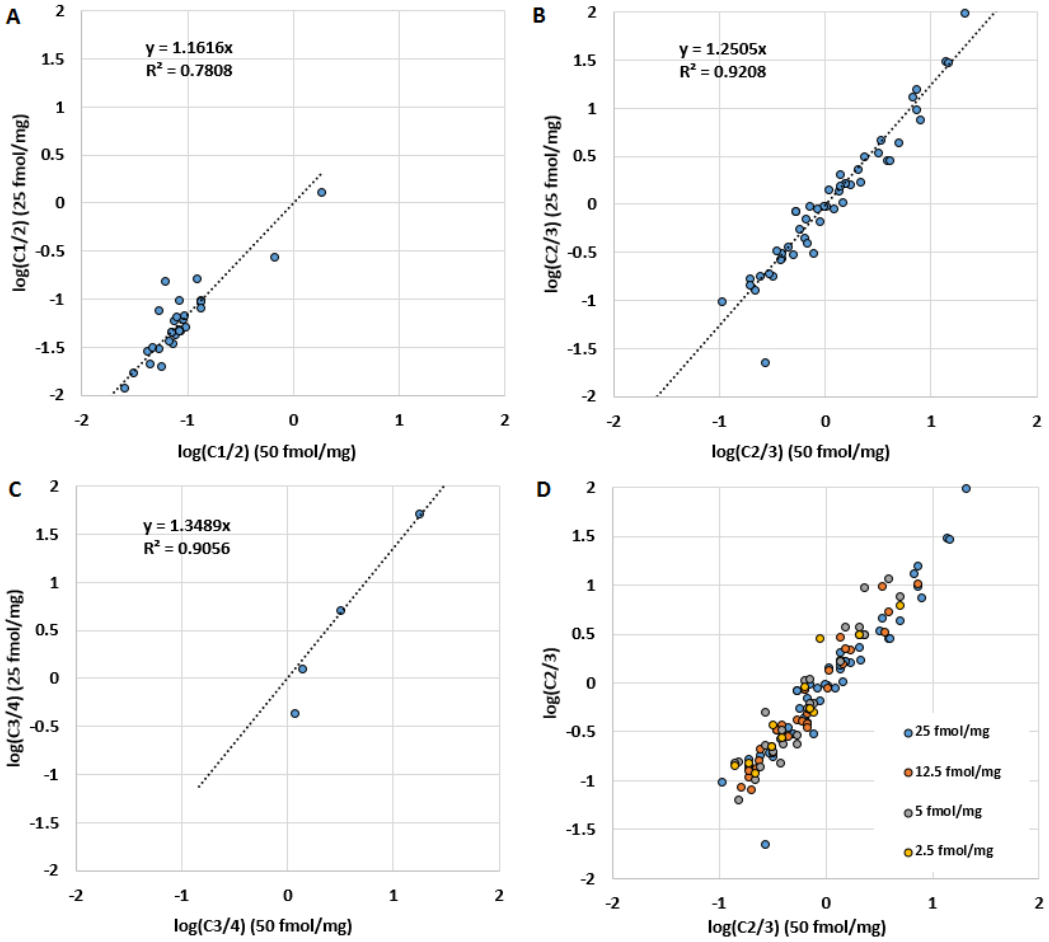

Рисунок 3. Примеры сравнения величины $C_{n/m}$ (соотношение между суммами интенсивностей ионов разного заряда), полученной для различных разведений.

В первом рассматривали долю варианта иона пептида конкретной зарядности (ω_{n+}), вычисленную как отношение суммы интенсивностей данного иона к сумме всех сумм интенсивностей обнаруженных ионов этого пептида. Для любого пептида сумма всех величин ω_{n+} равна 1. В данной работе рассматривали зарядность ионов от 1+ до 5+ (рис. 1). Были зафиксированы единичные случаи наличия ионов 6+, но они не рассматривались. К сожалению, определить достоверно долю пептидов, не имеющих положительного заряда из эксперимента нельзя. Второй набор величин характеризует соотношение между суммами интенсивностей ионов ($C_{n/m}$). В данной работе рассмотрены только 3 пары ионов 1+/2+, 2+/3+ и 3+/4+. Величина $C_{n/m}$ имеет смысл только в случае существования обоих ионов. Проблема существует и для величины ω_{n+}, если она принимает значение 0 или 1. Нет уверенности, что пептид в подавляющем количестве существует в виде данного иона, или иона не существует вовсе. Возможно этот ион или другие ионы пептида просто не были зарегистрированы, либо неправильно идентифицированы программным обеспечением.

Так как распределение ионов по заряду для отдельного пептида определяется его химической природой (или аминокислотной последовательностью), то оно не должно зависеть от концентрации пептида. Рисунок 2 демонстрирует подтверждение этого постулата. Как и предполагалось изначально, можно наблюдать критические отклонения в случаях, когда величина ω_{n+} равна 0 или 1. Если отбросить эти значения, то корреляция становится более чем убедительная (табл. 1). Значения $C_{n/m}$ также не меняются от разведения к разведению (рис. 3, табл. 1). В последнем случае значения для пептидов, соответствующие ω_{n+}, равному 0 или 1, отсутствовали по определению. Таким образом, данные для величин ω_{n+} и $C_{n/m}$ можно объединить в одну выборку. Так как наиболее полная выборка - для разведения 50 fmol/mg, то её использовали как базовую, добавляя значения в случае, если пептид обнаруживался только при большем разведении, или заменяя значения в случае, если при большем разведении обнаруживали больше вариантов иона пептида. Если изначально в выборке для разведения 50 fmol/mg было 258 наблюдений с величинами ω_{2+}, отличными от 0 или 1, то в обобщённой выборке таких наблюдения стало 300.

Для создания набора уравнений, с помощью которых можно предсказать распределение ионов по заряду в качестве независимых переменных, использовали аминокислотный спектр пептида (количество каждого из 20 аминокислотных остатков, присутствующих в пептиде). В работе рассматриваются уравнения линейной регрессии. Для данного случая это не оптимальный выбор, так как при использовании линейной регрессии нельзя учесть тот факт, что количественные доли связаны между собой и дают в сумме единицу, а также, что в общем случае не вводятся граничные условия о том, что предсказываемая величина может принимать значения только от 0 до 1. Однако для демонстрации принципиальной возможности такого предсказания линейные уравнения использовать можно. Результаты представлены на рисунке 4 и в таблице 1. Видно, что линейная регрессия даёт удовлетворительный результат во всех случаях, когда количество наблюдений достаточно.

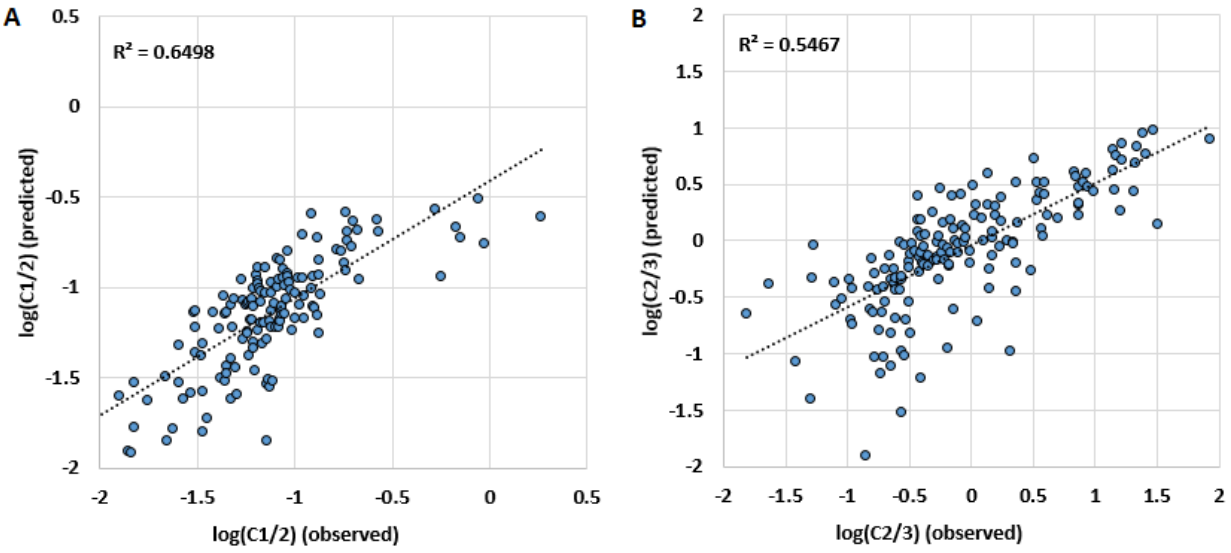

Рисунок 4. Сравнение наблюдаемых и предсказанных величин Cn/m в обучении.

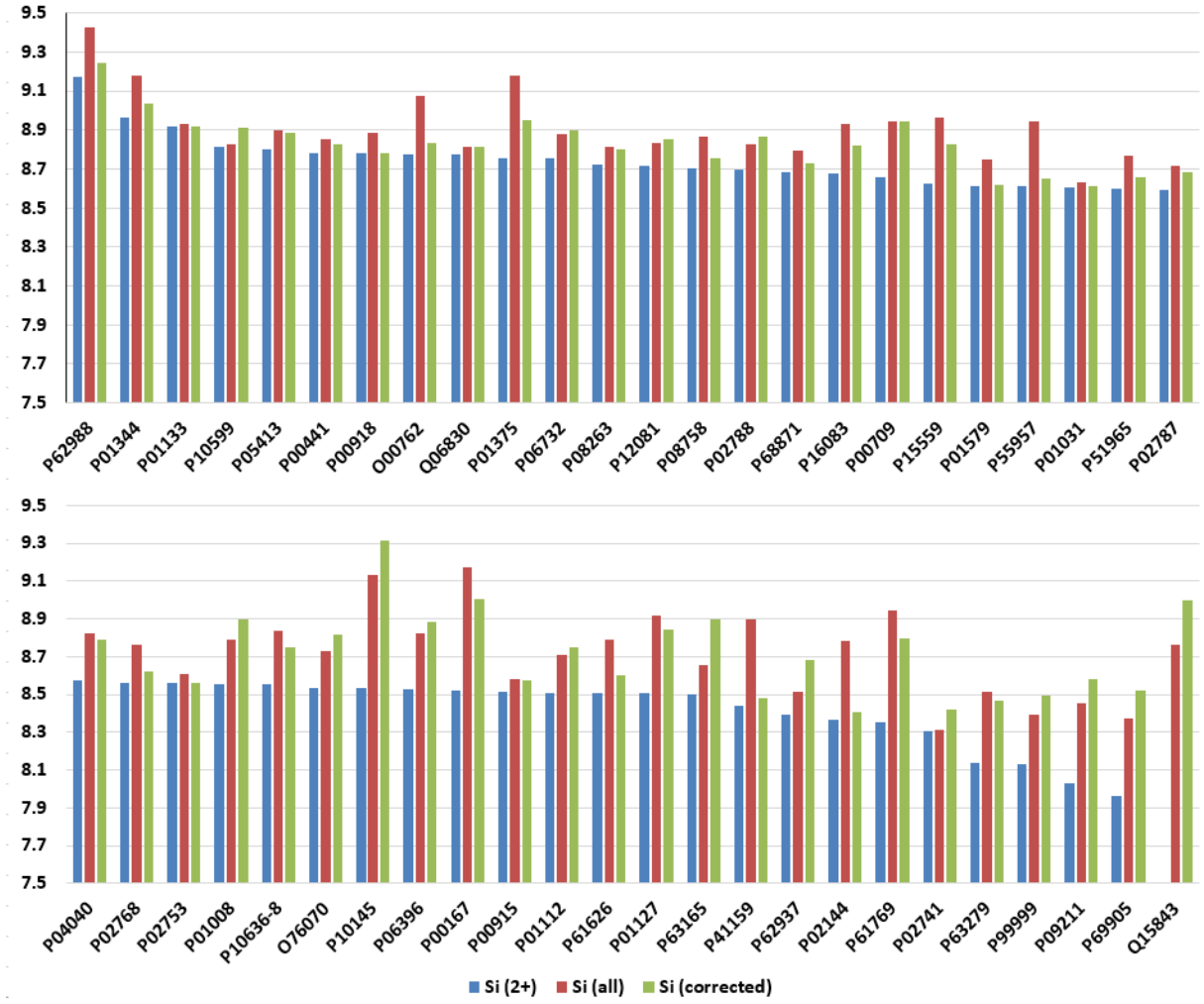

Рисунок 5. Пример коррекции значения суммарной интенсивности с учётом предсказания величин C1/2 и C2/3 (логарифмическая шкала).

Наилучший результат показан для величины $C_{n/m}$, особенно в случае $C_{1/2}$. Больший разброс для величины $C_{2/3}$, вероятно, связан с тем, что не учитывается наличие ионов с большим зарядом чем 3+, в то время как для $C_{1/2}$ большая часть наблюдений имеет максимально возможный заряд 2+.

Продемонстрируем пользу такого рода предсказания на примере разведения 50 fmol/mg. На рисунке 5 представлена усреднённая величина S_i для каждого из 48 белков набора UPS1. В первом случае были учтены только ионы 2+, наблюдаемые в эксперименте. При этом разброс

величины S_i лежит в пределах 1.21 логарифмической шкалы, а для одного из белков (Q15843) ионы 2+ не были зарегистрированы. Напомним, что количество каждого белка в наборе UPS1 одинаковое, а усреднение для каждого из них производили, за редким исключением, по четырём и больше пептидам (максимально 28), что должно нивелировать возможные различия, связанные с эффективностью трипсинолиза. Ожидаемым результатом было бы примерное равенство величин S_i. Если учитывать суммы S_i для всех зарегистрированных ионов, то разброс незначительно уменьшается до 1.12 логарифмической шкалы. Однако если провести коррекцию с использованием предсказания величин $C_{1/2}$ и $C_{2/3}$ для наблюдений, в которых были обнаружены только ионы 2+, и величины $C_{2/3}$, когда обнаруживались исключительно ионы 3+, то разброс уменьшается до 0.9 логарифмической шкалы. При этом эти пептиды заведомо не использовались в обучающей выборке, и, повторимся, линейная регрессия – не лучший метод для решения данной задачи. Тем не менее, даже при использовании очень простых подходов предсказать *a priory* долю пептида, приходящуюся на ион заданного заряда при ESI, можно.

ФИНАНСИРОВАНИЕ

Работа выполнена в рамках государственного задания по программе фундаментальных исследований Государственных академий.

ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ

К данной статье приложены дополнительные материалы, свободно доступные в электронной версии (http://dx.doi.org/10.18097/BMCRM00100) на сайте журнала.

ЛИТЕРАТУРА

1. Ya*tes, J. R., Ruse, C. I., & Nakorchevsky, A.* (2009). Proteomics by mass spectrometry: approaches, advances, and applications. Annual review of biomedical engineering, **11**, 49-79.
DOI: 10.1146/annurev-bioeng-061008-124934

2. Iavarone, A. T., Jurchen, J. C., & Williams, E. R. (2000). Effects of solvent on the maximum charge state and charge state distribution of protein ions produced by electrospray ionization. Journal of the American Society for Mass Spectrometry, **11**(11), 976-985. DOI: 10.1016/S1044-0305(00)00169-0

3. *Ramus, C., Hovasse, A., Marcellin, M., Hesse, A. M., Mouton-Barbosa, E., Bouyssié, D., ... & Garin, J.* (2016). Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. Data in brief, **6**, 286-294. DOI: 10.1016/j.dib.2015.11.063

4. https://www.ebi.ac.uk/pride, Submission Reference: PXD001819.

5. Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nature biotechnology, **26**(12), 1367. DOI: 10.1038/nbt.1511

6. *Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M.* (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. Journal of proteome research, **10**(4), 1794-1805.
DOI: 10.1021/pr101065j

7. Teleman, J., Chawade, A., Sandin, M., Levander, F., & Malmström, J. (2016). Dinosaur: a refined open-source peptide MS feature detector. Journal of proteome research, **15**(7), 2143-2151. DOI: 10.1021/acs.jproteome.6b00016

Поступила: 10.07.2019
После доработки: 10.10.2019
Принята к публикации: 14.10.2019

THE PREDICTION OF THE ION FRACTION OF THE PEPTIDE WITH SELECTED CHARGE IN MASS SPECTROMETRY WITH POSITIVE ELECTROSPRAY IONIZATION

V.S. Skvortsov *, N.N. Alekseychuk, Yu.V. Miroshnichenko, A.V. Rybina

Institute of Biomedical Chemistry, 10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: vladlen@ibmh.msk.su

The possibility of prediction of selected ion fraction in the total peptide fraction obtained during mass spectrometry with positive ionization by electrospray was investigated on the basis of the amino acid sequence. The data obtained in the MS / MS experiment [Ramus et al., 2015] using the standardized UPS1 kit (48 highly purified human proteins) and deposited in ProteomeXchange (identifier PXD001819) were used as the initial data set. For each of the identified peptides belonging to one of the proteins of the UPS kit, a list of detected ions of different charge was formed. The sum of the peak intensities detected for the primary ion was used as a measure of quantity. Since the ratio of the peptide fractions of ions with different charges does not depend on the concentration in the experimental sample, the total sample was assembled by combining the data obtained for different dilutions of UPS1. A set of equations of prediction of the fraction of 1+, 2+, and 3+ ions has been constructed. This computational analysis has shown applicability of the proposed for prediction of the ion fraction of the peptide with selected charge in mass spectrometry with positive electrospray ionization.

Key words: peptide; mass-spectrometry; electrospray ionization; property prediction

FUNDING

This work was performed within the framework of the Program for Basic Research of State Academies of Sciences for 2013-2020.

SUPPLEMENTARY

Supplementary materials are available at http://dx.doi.org/10.18097/BMCRM00100

Received: 10.-7.2019, revised: 10.10.2019, accepted: 14.10.2019